

# Разбор летучки

---

# Лекция 11

Анализ смещения и разброса

---

Екатерина Тузова

# Постановка задачи

$X$  – множество объектов

$Y$  – ответы в  $\mathbb{R}$

Обучающая выборка:  $X^l = (x_i, y_i)_{i=1}^l$

Целевая функция:  $f : X \rightarrow Y$

Набор моделей алгоритмов  $A_t : X \rightarrow Y, t \in T$

Методы обучения  $\mu : (X \times Y)^l \rightarrow A_t, t \in T$

$$E_{in} = Q_{\mu}(X^l) = Q(\mu(X^l), X^l)$$

Внешний функционал по отложенной выборке:

$$E_{out} = Q_{\mu}(X^t, X^k) = Q(\mu(X^t), X^k)$$

$$E_{in} = Q(w, X^l) = \sum_{i=1}^l (f(x_i, w) - y_i)^2 = \|Xw - y\|^2 \rightarrow \min_w$$

Внешний функционал по отложенной выборке:

$$E_{out} = Q(w^*, X^k) = \|X^k w^* - y\|^2$$

# Постановка задачи

Хотим:  $E_{out} \rightarrow \min$

Хотим:  $E_{out} \rightarrow \min$

$|A_t| \uparrow \Rightarrow$  больше шансов, что целевая функция  $f$  находится во множестве

$|A_t| \downarrow \Rightarrow$  лучше обобщающая способность алгоритма

Ошибка алгоритма складывается из двух частей:

- Насколько хорошо  $A_t$  может приблизить целевую функцию  $f$
- Можем ли мы на основании  $X^l$  выбрать  $f$  из  $A_t$



Квадратичная функция потерь:

$$L_i(a) = (a(x_i) - f(x_i))^2$$

Внутренний функционал:

$$E_{in} = \sum_{i=1}^l (a(x_i, w) - f(x_i))^2 \rightarrow \min_w$$

$$E_{out}(a^{(X^l)}) = \mathbb{E}_{\mathbf{x}} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right]$$

$$E_{out}(a^{(X^l)}) = \mathbb{E}_{\mathbf{x}} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right]$$

$$\begin{aligned} \mathbb{E}_{(X^l)} \left[ E_{out}(a^{(X^l)}) \right] &= \mathbb{E}_{(X^l)} \left[ \mathbb{E}_{\mathbf{x}} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right] \right] \\ &= \mathbb{E}_{\mathbf{x}} \left[ \mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right] \right] \end{aligned}$$

$$\mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right]$$

$$\mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right]$$

$$\bar{a}(\mathbf{x}) = \mathbb{E}_{(X^l)} \left[ a^{(X^l)}(\mathbf{x}) \right] \approx \frac{1}{K} \sum_{k=1}^K a^{X_k}(\mathbf{x})$$

$X_1, \dots, X_K$  – различные обучающие выборки

$$\mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right] = \mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}) + \bar{a}(\mathbf{x}) - f(\mathbf{x}))^2 \right]$$

$$\begin{aligned}\mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right] &= \mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}) + \bar{a}(\mathbf{x}) - f(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))^2 + (\bar{a}(\mathbf{x}) - f(\mathbf{x}))^2 \right. \\ &\quad \left. + 2(a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))(\bar{a}(\mathbf{x}) - f(\mathbf{x})) \right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2 \right] &= \mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}) + \bar{a}(\mathbf{x}) - f(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))^2 + (\bar{a}(\mathbf{x}) - f(\mathbf{x}))^2 \right. \\ &\quad \left. + 2(a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))(\bar{a}(\mathbf{x}) - f(\mathbf{x})) \right] \\ &= \mathbb{E}_{(X^l)} \left[ (a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))^2 \right] + (\bar{a}(\mathbf{x}) - f(\mathbf{x}))^2\end{aligned}$$



$$\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2] = \underbrace{\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))^2]}_{\text{var}(\mathbf{x})} + \underbrace{(\bar{a}(\mathbf{x}) - f(\mathbf{x}))^2}_{\text{bias}(\mathbf{x})}$$

$$\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2] = \underbrace{\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))^2]}_{\text{var}(\mathbf{x})} + \underbrace{(\bar{a}(\mathbf{x}) - f(\mathbf{x}))^2}_{\text{bias}(\mathbf{x})}$$

$$\mathbb{E}_{(X^l)}[E_{out}(a^{(X^l)})] = \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2]]$$

$$\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2] = \underbrace{\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))^2]}_{var(\mathbf{x})} + \underbrace{(\bar{a}(\mathbf{x}) - f(\mathbf{x}))^2}_{bias(\mathbf{x})}$$

$$\begin{aligned}\mathbb{E}_{(X^l)}[E_{out}(a^{(X^l)})] &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2]] \\ &= \mathbb{E}_{\mathbf{x}}[var(\mathbf{x}) + bias(\mathbf{x})]\end{aligned}$$

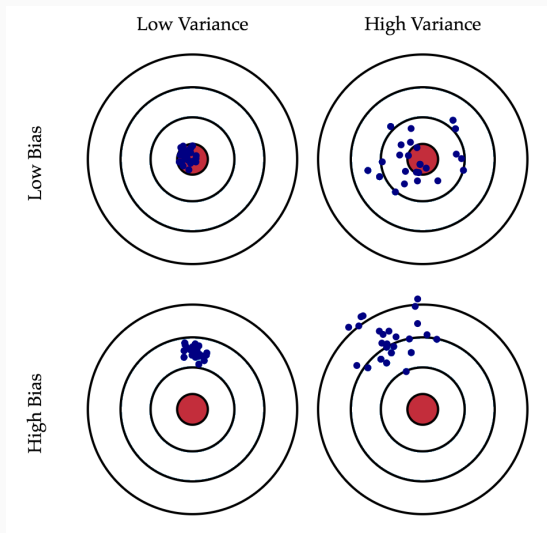
$$\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2] = \underbrace{\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - \bar{a}(\mathbf{x}))^2]}_{var(\mathbf{x})} + \underbrace{(\bar{a}(\mathbf{x}) - f(\mathbf{x}))^2}_{bias(\mathbf{x})}$$

$$\begin{aligned}\mathbb{E}_{(X^l)}[E_{out}(a^{(X^l)})] &= \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{(X^l)}[(a^{(X^l)}(\mathbf{x}) - f(\mathbf{x}))^2]] \\ &= \mathbb{E}_{\mathbf{x}}[var(\mathbf{x}) + bias(\mathbf{x})] \\ &= var + bias\end{aligned}$$

Смещение (bias) – насколько сложное семейство моделей (средняя ошибка по всевозможным обучающим выборкам  $X^l$ )

Разброс (variance) – на сколько чувствителен алгоритм к изменению обучающей выборки (как отличается ошибка, если обучать модель на разных наборах данных)

# Смещение и разброс



Как правило, при увеличении сложности модели увеличивается разброс оценки, но уменьшается смещение.

Как правило, при увеличении сложности модели увеличивается разброс оценки, но уменьшается смещение.

Если же модель слабая, то она не способна выучить закономерность, в результате выучивается что-то другое, смещенное относительно правильного решения.



# Пример $\sin$

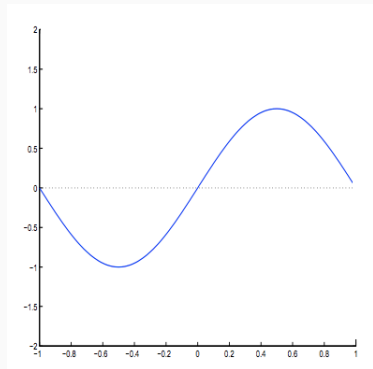
$$f(x) = \sin(\pi x)$$

Количество объектов  $l = 2$

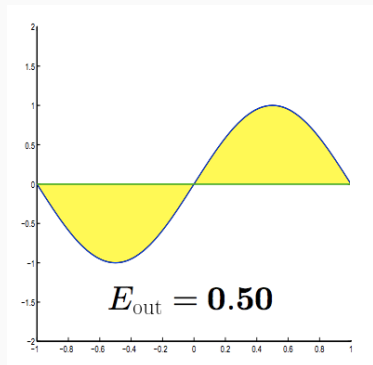
Рассмотрим 2 модели:

$$A_1 : h(x) = b$$

$$A_2 : h(x) = ax + b$$

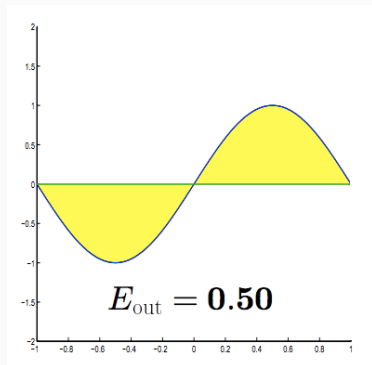


# Пример $\sin$

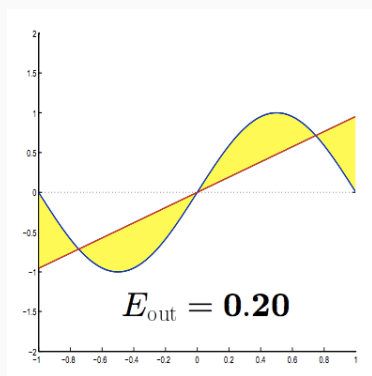


$A_1$

# Пример $\sin$

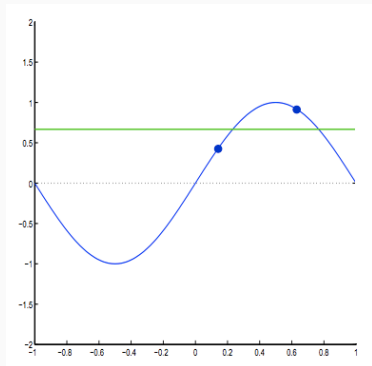


$A_1$



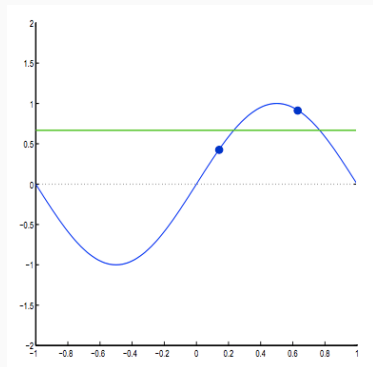
$A_2$

# Обучение $A_1, A_2$

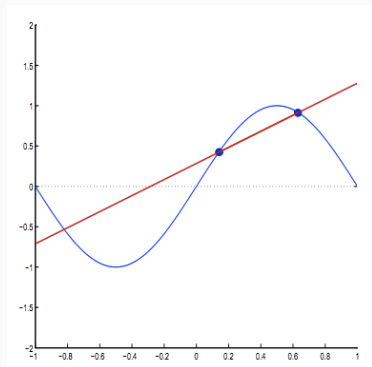


$A_1$

# Обучение $A_1, A_2$

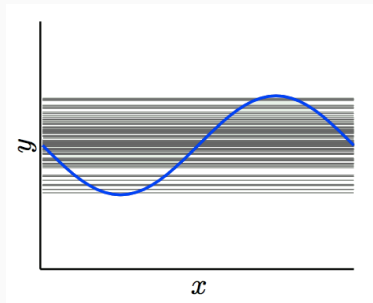


$A_1$

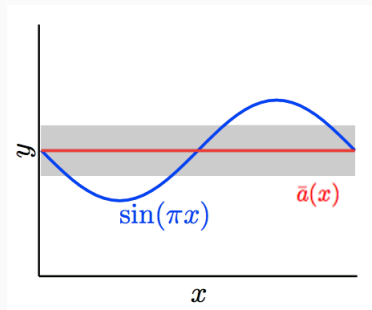
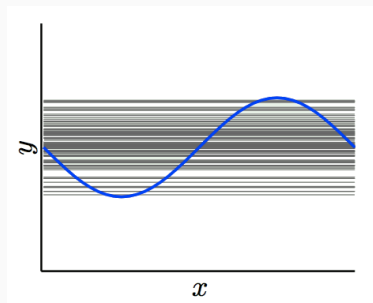


$A_2$

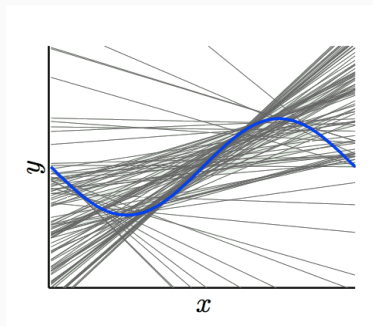
# Bias-variance $A_1$



# Bias-variance $A_1$

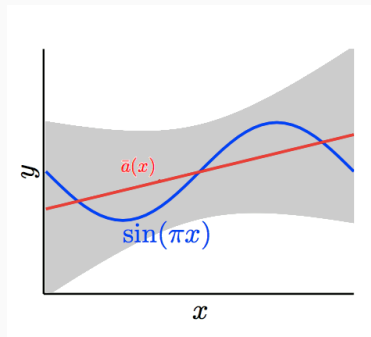
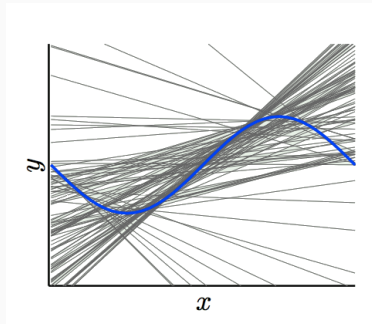


# Bias-variance $A_2$

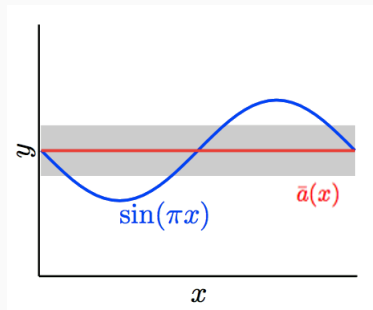




# Bias-variance $A_2$



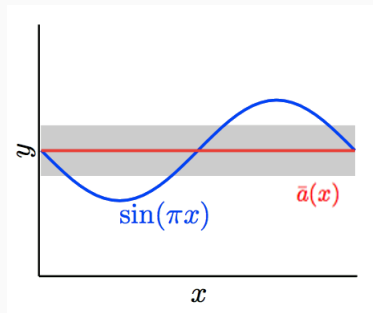
# Bias-variance



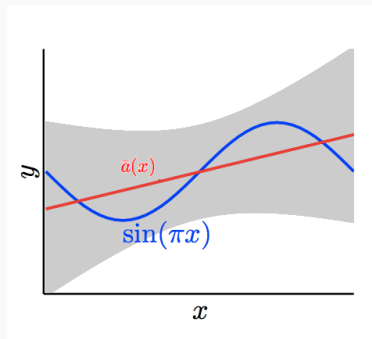
bias = 0.50

var = 0.25

# Bias-variance



bias = 0.50      var = 0.25



bias = 0.21      var = 1.69

Сложность модели нужно выбирать исходя из доступных данных, а не из предполагаемой сложности целевой функции.

# Кривые обучения

---

$$\mathbb{E}_{(X^l)} \left[ E_{out}(a^{(X^l)}) \right]$$

$$\mathbb{E}_{(X^l)} \left[ E_{out}(a^{(X^l)}) \right]$$

$$\mathbb{E}_{(X^l)} \left[ E_{in}(a^{(X^l)}) \right]$$

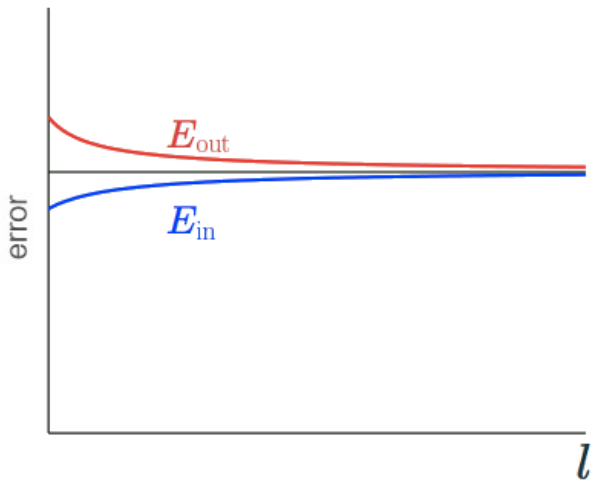
$$\mathbb{E}_{(X^l)} \left[ E_{out}(a^{(X^l)}) \right]$$

$$\mathbb{E}_{(X^l)} \left[ E_{in}(a^{(X^l)}) \right]$$

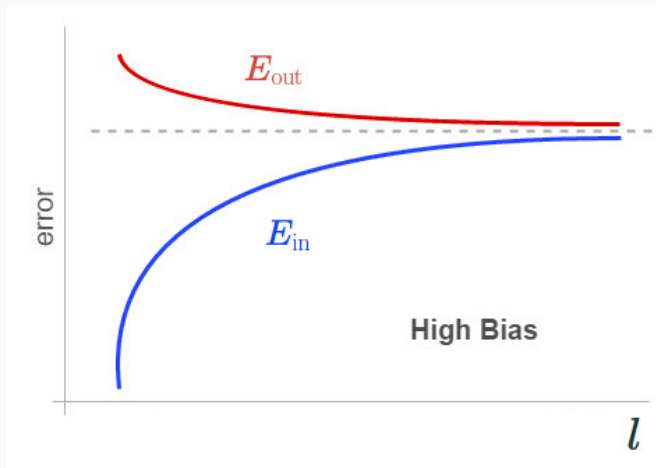
Как зависят от длины выборки?

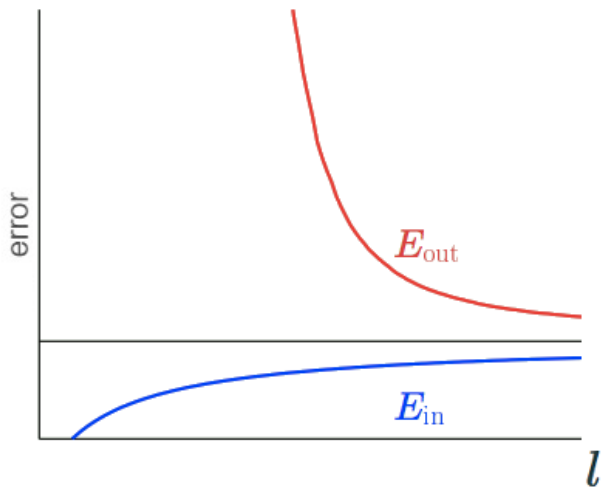


## Простая модель

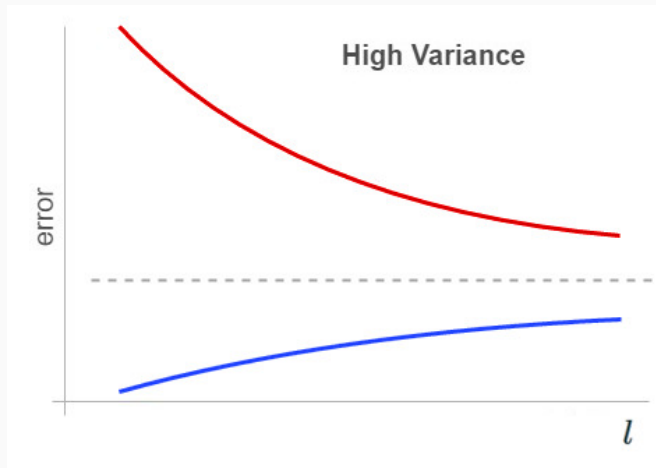


## High bias





# High variance



Вопросы?

## Что почитать по этой лекции

- Professor Yaser Abu-Mostafa MOOC
- Hastie, T., Tibshirani R. "The Elements of Statistical Learning"  
Chapter 7

## На следующей лекции

- Методы восстановления регрессии
- K neighbors regressor
- Decision tree regression
- Neural Net Regression
- SVR